

On the closure of relational models

Anna Klimova

Institute of Science and Technology (IST) Austria

aklimova25@gmail.com

Tamás Rudas

Eötvös Loránd University, Budapest, Hungary

rudas@tarki.hu

Abstract

Relational models for contingency tables are generalizations of log-linear models, allowing effects associated with arbitrary subsets of cells in a possibly incomplete table, and not necessarily containing the overall effect. In this generality, the MLEs under Poisson and multinomial sampling are not always identical. This paper deals with the theory of maximum likelihood estimation in the case when there are observed zeros in the data. A unique MLE to such data is shown to always exist in the set of pointwise limits of sequences of distributions in the original model. This set is equal to the closure of the original model with respect to the Bregman information divergence. The same variant of iterative scaling may be used to compute the MLE in the original model and in its closure.

Keywords: algebraic variety, Bregman divergence, contingency table, extended MLE, iterative scaling, relational model

1 Introduction

The existence of maximum likelihood estimates under log-linear models for contingency tables has been thoroughly studied, see Haberman [1974], Andersen [1974], Barndorff-Nielsen [1978], Lauritzen [1996], among others. It was established that the maximum likelihood estimates of the cell parameters always exist if the observed table has only positive cell counts, and may exist if some of the observed counts are zero. The patterns of zero cells that lead to the non-existence of the MLE were described in several forms [cf. Haberman, 1974, Fienberg and Rinaldo, 2012].

Within the extended log-linear model class all data sets have an MLE, irrespective of the pattern of zeros. An extended log-linear model may be obtained as the closure of the original model in the topology of pointwise convergence [cf. Lauritzen, 1996], or the closure with respect to the Kullback-Leibler divergence [cf. Csiszár and Matúš, 2003], or as the aggregate exponential family [Brown, 1988].

The contribution of this paper is motivated by statistical problems in which models more general than log-linear need to be considered. To illustrate, suppose that the management of a large supermarket classifies all goods on stock into one of three mutually exclusive and exhaustive categories, say, food (F), non-food household (N) and other (O), and wishes to study how the daily sales of each group are related. This is a standard task in market basket analysis [cf. Brin et al., 1997]. The first model of interest, routinely, is independence, but the usual model of independence of the three indicator variables is not applicable in this case: if p_F , p_N and p_O denote the probabilities that a purchase (a basket) contains an item from the F , N and O groups, then the probability of an empty purchase would be $(1 - p_F)(1 - p_N)(1 - p_O)$, which has to be positive, in spite of the fact that there are no purchases which do not contain any items.

One alternative independence concept to apply is the AS-independence of the three variables [Aitchison and Silvey, 1960]. The indicator variables F , N , and O are said to be AS-independent if

$$p_{FN} = p_F p_N, \quad p_{FO} = p_F p_O, \quad p_{NO} = p_N p_O, \quad p_{FNO} = p_F p_N p_O. \quad (1)$$

Relational models introduced by Klimova, Rudas, and Dobra [2012] contain model (1) and many other models of association.

A relational model on a contingency table is generated by a class of non-empty subsets of cells and can be specified in the form:

$$\log \boldsymbol{\delta} = \mathbf{A}'\boldsymbol{\beta}. \quad (2)$$

Here, $\boldsymbol{\delta}$ denotes the vector of cell parameters, probabilities or intensities, and \mathbf{A} is the 0-1 matrix whose rows are the indicators of generating subsets. A hierarchical log-linear model [cf. Bishop, Fienberg, and Holland, 1975] applies to a table which is a Cartesian product, and the model is generated by a collection of cylinder sets corresponding to marginals of the table and thus is a special case of a relational model. If the row space of \mathbf{A} contains the vector $\mathbf{1}' = (1, \dots, 1)$, as in the case of hierarchical log-linear models, then the model is said to include the overall effect. A model with the overall effect can be parameterized to include a common parameter in every cell, often called the normalizing constant. The models without the overall effect cannot be parameterized in such a way. The peculiar property of relational models without the overall effect is that models for probabilities (appropriate under multinomial sampling) and models for intensities (appropriate for Poisson sampling) are different and lead to different MLEs. Let \mathbf{y} denote the observed frequency distribution. Then, when the overall effect is not present, the MLE for probabilities does not preserve the sufficient statistics $\mathbf{A}\mathbf{y}$, and, for intensities, it does not preserve the observed total $\mathbf{1}'\mathbf{y}$, see Example 2.1.

An iterative scaling procedure based on Bregman divergence can be used to compute the MLE under relational models [Klimova and Rudas, 2015]. The Bregman divergence between two distributions is a generalization of the Kullback-Leibler divergence, but, unlike the latter, stays non-negative whether or not the two distributions have the same total. This property is essential for relational models for intensities without the overall effect as these models may include distributions with different totals.

If the observed frequencies are positive and the model matrix is of full row rank, the MLE under relational models can be computed using algorithms for convex optimization

[cf. Bertsekas, 1999, Aitchison and Silvey, 1960, Evans and Forcina, 2013] or the Newton-Raphson algorithm. A detailed discussion of the relative advantages and disadvantages of variants of iterative proportional fitting was given in Klimova and Rudas [2015]. The contribution of the present paper is the investigation of cases when there are observed zero frequencies in the data, and of the closure of relational models under which such data will always admit an MLE. Of course, if only three groups of goods, as in the example above, are investigated, one cannot expect to see an observed zero, but if 1000 groups of goods are investigated, out of the resulting $2^{1000} - 1$ groups, many will be empty. As it turns out, the pattern of observed zeros has far reaching implications on the existence and kind of MLE obtained.

A necessary and sufficient condition for the existence of the maximum likelihood estimates of the cell parameters under relational models is obtained in Section 2. The MLE for \mathbf{y} exists if and only if there is a positive vector \mathbf{z} such that $\mathbf{Az} = \mathbf{Ay}$. This is literally the same condition as the one that applies to log-linear models.

In Section 3, extended relational models are studied. The extended relational model is defined as the set of distributions parameterized by the elements of an algebraic variety associated with the model matrix of the original relational model. It is shown that this set is equal to the closure of the original model with respect to both the pointwise convergence and the Bregman divergence.

In Section 4, a polyhedral condition for the existence of the MLE in the original or the extended relational model is formulated. If the vector of the sufficient statistics, \mathbf{Ay} , of the observed distribution is not contained in any of the faces of the polyhedral cone associated with the model matrix, the MLE exists in the original model, and otherwise, it does in the extended model. This condition is the same as for the log-linear case, but the proof is very different. The multiplicative representation of the distributions in the extended model and the existence of the MLEs of the model parameters are also discussed in this section. Finally, the generalized iterative proportional fitting procedure suggested in Klimova and Rudas [2015] is extended to the case of observed zeros.

While the conditions of the existence of the MLE in the generality considered in this paper may be formulated to coincide with the known conditions for the case of log-linear models, the proofs turn out to be more involved. Also, the algorithm to obtain that the MLEs is more complex. The additional complications come from properties of the MLE when the overall effect is not present. In fact, Lauritzen [1996, p.75] mentioned the existence of models without the overall effect, which he called the “constant function”, but to avoid difficulties did not consider them. On the other hand, such models have been used in practice, see references in Klimova et al. [2012], Klimova and Rudas [2015].

2 MLE under relational models

Let Y_1, \dots, Y_K be discrete random variables with finite ranges, and the vector \mathcal{I} of length $|\mathcal{I}|$ be their joint sample space. Here, \mathcal{I} may also be a proper subset of the Cartesian product of the ranges of the variables. A distribution on \mathcal{I} is parameterized by the cell parameters $\boldsymbol{\delta} = \{\delta_i, \text{ for } i \in \mathcal{I}\}$, and, to simplify notation, is identified with $\boldsymbol{\delta}$. The components of $\boldsymbol{\delta}$ are either probabilities: $\delta_i \equiv p_i \in (0, 1)$, with $\sum_{i \in \mathcal{I}} p_i = 1$, or intensities: $\delta_i \equiv \lambda_i > 0$, for all

$i \in \mathcal{I}$. Let \mathcal{P} denote the set of positive distributions, $\boldsymbol{\delta} > \mathbf{0}$, on \mathcal{I} .

Let \mathbf{A} be a 0-1 matrix of size $J \times |\mathcal{I}|$, which is interpreted as the indicator matrix of J subsets generating the model. Assume that \mathbf{A} has no zero column. A relational model $RM_{\boldsymbol{\delta}}(\mathbf{A})$ is the following set of distributions:

$$RM_{\boldsymbol{\delta}}(\mathbf{A}) = \{\boldsymbol{\delta} \in \mathcal{P} : \delta_i = \prod_{j=1}^J \theta_j^{a_{ji}}, i \in \mathcal{I}, \text{ for some } \boldsymbol{\theta} \in \mathbb{R}_{>0}^J\}, \quad (3)$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J) \in \mathbb{R}_{>0}^J$ denotes the vector of parameters associated with the generating subsets. Under the model, the cell parameters are equal to the products of the parameters $\boldsymbol{\theta}$ corresponding to the subsets to which the cell belongs. In the sequel, the components of $\boldsymbol{\theta}$ are referred to as the multiplicative parameters, and \mathbf{A} is assumed to be of full row rank. In fact, the model $RM_{\boldsymbol{\delta}}(\mathbf{A})$ is uniquely determined by the row space of its model matrix, $\mathcal{R}(\mathbf{A})$. Relational models for which $\mathbf{1}' \in \mathcal{R}(\mathbf{A})$ are said to include the overall effect.

A dual representation of a relational model $RM_{\boldsymbol{\delta}}(\mathbf{A})$ can be obtained using the kernel basis matrix \mathbf{D} , whose rows, $\mathbf{d}_1, \dots, \mathbf{d}_K$, are a basis of $Ker(\mathbf{A})$. In this representation, any distribution in the model satisfies

$$\mathbf{D} \log \boldsymbol{\delta} = \mathbf{0}, \quad (4)$$

which can be re-written using the generalized odds ratios:

$$\delta^{d_1^+} / \delta^{d_1^-} = 1, \quad \delta^{d_2^+} / \delta^{d_2^-} = 1, \quad \dots \quad \delta^{d_K^+} / \delta^{d_K^-} = 1, \quad (5)$$

or using the cross-product differences:

$$\delta^{d_1^+} - \delta^{d_1^-} = 0, \quad \delta^{d_2^+} - \delta^{d_2^-} = 0, \quad \dots \quad \delta^{d_K^+} - \delta^{d_K^-} = 0, \quad (6)$$

where, \mathbf{d}^+ and \mathbf{d}^- denote, respectively, the positive and negative parts of a vector \mathbf{d} [Klimova et al., 2012].

The properties of the maximum likelihood estimates under relational models are reviewed next. Let $\mathbf{Y} = (Y_1, \dots, Y_K)$ be a random variable that has a multivariate Poisson distribution parameterized by $\boldsymbol{\delta} \equiv \boldsymbol{\lambda}$ or a multinomial distribution parameterized by N and $\boldsymbol{\delta} \equiv \mathbf{p}$. Let \mathbf{y} be a realization of \mathbf{Y} , and

$$\mathbf{q} = \begin{cases} \mathbf{y}, & \text{if } \boldsymbol{\delta} \equiv \boldsymbol{\lambda}, \\ \mathbf{y}/(\mathbf{1}'\mathbf{y}), & \text{if } \boldsymbol{\delta} \equiv \mathbf{p}. \end{cases} \quad (7)$$

If the MLE $\hat{\boldsymbol{\delta}}_{\mathbf{y}}$ of the cell parameters under the model $RM_{\boldsymbol{\delta}}(\mathbf{A})$ exists, it is the unique solution to the system of equations:

$$\begin{aligned} \mathbf{A}\boldsymbol{\delta} &= \gamma \mathbf{A}\mathbf{q}, \\ \mathbf{D} \log \boldsymbol{\delta} &= \mathbf{0}, \\ \mathbf{1}'\boldsymbol{\delta} &= 1 \quad (\text{only for } \boldsymbol{\delta} \equiv \mathbf{p}). \end{aligned} \quad (8)$$

The value of γ is called the adjustment factor. If $RM_{\boldsymbol{\delta}}(\mathbf{A})$ is a model for probabilities with the overall effect or a model for intensities, then $\gamma = 1$ for every \mathbf{y} . If $RM_{\boldsymbol{\delta}}(\mathbf{A})$ is a model for probabilities without the overall effect, then the value of γ depends on \mathbf{y} [Klimova et al., 2012].

Table 1: Maximum likelihood estimates under the model of AS-independence of variables F , N , O , under the multinomial and Poisson sampling.

		$O = No$		$O = Yes$	
		$N = No$	$N = Yes$	$N = No$	$N = Yes$
$F = No$	<i>empty</i>	14	25	16	- <i>observed</i>
	<i>empty</i>	27.33	32.60	8.91	- <i>multinomial</i>
	<i>empty</i>	3.31	7.29	24.13	- <i>Poisson</i>
$F = Yes$	10	5	3	27	- <i>observed</i>
	18.46	5.04	6.02	1.64	- <i>multinomial</i>
	1.26	4.17	9.18	30.39	- <i>Poisson</i>

Example 2.1. The model of AS-independence (1) is a relational model generated by the model matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}, \quad (9)$$

where the order of cells is lexicographic. As $\mathbf{1}'$ is not in the row space of \mathbf{A} , the model does not have the overall effect. Thus, the models $RM_{\lambda}(\mathbf{A})$ and $RM_p(\mathbf{A})$ are not equivalent. Given hypothetical data, the MLE for cell frequencies, computed under the model for probabilities and under the model for intensities, are shown in Table 1. In the case of probabilities, the estimates for sufficient statistics are about 0.7 times less than the observed sufficient statistics. In the case of intensities, the estimated total is approximately 79.73, while the observed total is 100. The estimates were obtained using the R-package `gIPFrm` [Klimova and Rudas, 2014]. \square

A necessary and sufficient condition for the existence of the MLE is given in the next theorem. Its proof uses the following lemma:

Lemma 2.1. *If $\mathbf{y} > \mathbf{0}$, the MLE $\hat{\boldsymbol{\delta}}_{\mathbf{y}}$ exists.*

Proof. A relational model for intensities is a regular exponential family [Klimova et al., 2012], and the standard proof applies [cf. Andersen, 1974].

In the case of probabilities, $\boldsymbol{\delta} \equiv \mathbf{p}$, the MLE, if exists, is the unique solution to (8). Klimova and Rudas [2015, Lemma 3.5] showed that there exist $\gamma_1, \gamma_2 > 0$ such that the adjustment factor $\gamma \in [\gamma_1, \gamma_2]$. Since $\gamma \mathbf{y} > \mathbf{0}$, the MLE $\hat{\boldsymbol{\lambda}}_{\gamma \mathbf{y}}$ under the model for intensities $RM_{\lambda}(\mathbf{A})$ exists for every $\gamma \in [\gamma_1, \gamma_2]$, and, by Lemma 3.6 in Klimova and Rudas [2015], one can find a unique γ^* such that $\mathbf{1}' \hat{\boldsymbol{\lambda}}_{\gamma^* \mathbf{y}} = 1$. Because $\hat{\boldsymbol{\lambda}}_{\gamma^* \mathbf{y}}$ satisfies (8), $\hat{\mathbf{p}}_{\mathbf{y}} = \hat{\boldsymbol{\lambda}}_{\gamma^* \mathbf{y}}$. \square

As shown next, the MLE may exists when some of the observed frequencies are zero.

Example 2.1 (revisited):

Let $\mathbf{q} = (0, 0, 0, 0, 0, 0, 1)'$ be the observed distribution. Under the model of AS-independence, the MLEs for cell probabilities exist and are equal to

$$\hat{\mathbf{p}} = \left(\sqrt[3]{2} - 1, \sqrt[3]{2} - 1, \sqrt[3]{2} - 1, (\sqrt[3]{2} - 1)^2, (\sqrt[3]{2} - 1)^2, (\sqrt[3]{2} - 1)^2, (\sqrt[3]{2} - 1)^3 \right)'.$$

□

Theorem 2.2. *Let \mathbf{y} be the vector of observed frequencies under Poisson or multinomial sampling, and let $RM_{\delta}(\mathbf{A})$ be a relational model. The MLE $\hat{\delta}_{\mathbf{y}}$ under the model exists if and only if there is a positive vector \mathbf{z} , such that $\mathbf{Az} = \mathbf{Aq}$, with \mathbf{q} defined in (7).*

Proof. In the case of intensities, $\delta \equiv \lambda$, the standard proof for regular exponential families [cf. Andersen, 1974] applies.

The case of probabilities, $\delta \equiv \mathbf{p}$, is considered next. Suppose $\hat{\mathbf{p}}_{\mathbf{y}} > 0$ exists. By Corollary 4.2 in Klimova et al. [2012], $\mathbf{A}\hat{\mathbf{p}}_{\mathbf{y}} = \gamma\mathbf{Aq}$ for some $\gamma > 0$. Therefore, $\hat{\mathbf{p}}_{\mathbf{y}} = \gamma\mathbf{q} + \mathbf{d}$, for some $\mathbf{d} \in \text{Ker}(\mathbf{A})$. Take

$$\mathbf{z} = \frac{1}{\gamma}\hat{\mathbf{p}}_{\mathbf{y}} = \mathbf{q} + \frac{1}{\gamma}\mathbf{d} > \mathbf{0}.$$

Then, as $\frac{1}{\gamma}\mathbf{d} \in \text{Ker}(\mathbf{A})$, $\mathbf{Az} = \mathbf{Aq}$, as required.

To prove the converse, assume that there exists a $\mathbf{z} > \mathbf{0}$, such that $\mathbf{Az} = \mathbf{Aq}$. Thus, $\mathbf{z} = \mathbf{q} + \mathbf{d}$ for some $\mathbf{d} \in \text{Ker}(\mathbf{A})$. Let

$$\mathbf{d}_1 = \frac{1}{1 + \mathbf{1}'\mathbf{d}}\mathbf{d},$$

and note that $1 + \mathbf{1}'\mathbf{d} = \mathbf{1}'\mathbf{q} + \mathbf{1}'\mathbf{d} = \mathbf{1}'\mathbf{z} > 0$. Next, consider $\mathbf{v} = (1 - \mathbf{1}'\mathbf{d}_1)\mathbf{q} + \mathbf{d}_1$. Then $\mathbf{1}'\mathbf{v} = (1 - \mathbf{1}'\mathbf{d}_1) + \mathbf{1}'\mathbf{d}_1 = 1$, and

$$\mathbf{v} = (1 - \mathbf{1}'\mathbf{d}_1)\mathbf{q} + \mathbf{d}_1 = \frac{1}{1 + \mathbf{1}'\mathbf{d}}\mathbf{q} + \frac{1}{1 + \mathbf{1}'\mathbf{d}}\mathbf{d} = \frac{1}{1 + \mathbf{1}'\mathbf{d}}(\mathbf{q} + \mathbf{d}) > \mathbf{0}.$$

Therefore, \mathbf{v} is a positive probability distribution, and, by Lemma 2.1, the MLE $\hat{\mathbf{p}}_{\mathbf{v}}$ exists, and satisfies:

$$\begin{aligned} \mathbf{A}\hat{\mathbf{p}}_{\mathbf{v}} &= \gamma_{\mathbf{v}}\mathbf{Av}, \\ \mathbf{D}\log \hat{\mathbf{p}}_{\mathbf{v}} &= \mathbf{0}, \\ \mathbf{1}'\hat{\mathbf{p}}_{\mathbf{v}} &= 1, \end{aligned}$$

for some $\gamma_{\mathbf{v}} > 0$. Then, from the definition of \mathbf{v} , $\mathbf{A}\hat{\mathbf{p}}_{\mathbf{v}} = \gamma_{\mathbf{v}}\mathbf{Av} = \gamma_{\mathbf{v}}(1 - \mathbf{1}'\mathbf{d}_1)\mathbf{Aq}$, that is, $\hat{\mathbf{p}}_{\mathbf{v}}$ is also the MLE for \mathbf{q} with the adjustment factor $\gamma = \gamma_{\mathbf{v}}(1 - \mathbf{1}'\mathbf{d}_1)$. □

The statement of the theorem is illustrated in the next example.

Example 2.2. Let $RM_{\mathbf{p}}(\mathbf{A})$ be the model for probabilities generated by

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix},$$

and let $\mathbf{q} = (3/7, 3/7, 0, 1/7, 0)'$ be the observed probability distribution. Consider any vector \mathbf{z} , whose subset sums, $\mathbf{A}\mathbf{z}$, are equal to the observed subset sums:

$$z_1 + z_2 + z_3 + z_5 = 6/7, \quad z_1 + z_2 + z_5 = 6/7, \quad z_1 + z_4 + z_5 = 4/7.$$

The first two equations imply that $z_3 = 0$. Therefore, there is no (strictly) positive distribution with the same subset sums as those observed, and thus, \mathbf{q} does not have an MLE in the model. \square

In the next section, an extended relational model is defined as the polynomial variety corresponding to the model matrix. It is further shown that the extended model coincides with the set of pointwise limits of sequences of distributions in the original model and is also the closure with respect to Bregman information divergence.

3 Extended relational models

Let \mathbf{A} be the model matrix of a relational model, and let $\mathcal{X}_{\mathbf{A}}$ denote the polynomial variety associated with \mathbf{A} [Sturmfels, 1996]:

$$\mathcal{X}_{\mathbf{A}} = \left\{ \boldsymbol{\delta} \in \mathbb{R}_{\geq 0}^{|\mathcal{I}|} : \boldsymbol{\delta}^{d^+} = \boldsymbol{\delta}^{d^-}, \forall d \in \text{Ker}(\mathbf{A}) \right\}. \quad (10)$$

Definition 3.1. The extended relational model for intensities, $\overline{RM}_{\boldsymbol{\lambda}}(\mathbf{A})$, is the set of distributions

$$\boldsymbol{\lambda} \in \mathcal{X}_{\mathbf{A}}. \quad (11)$$

The extended relational model for probabilities, $\overline{RM}_{\mathbf{p}}(\mathbf{A})$, is the set of distributions

$$\mathbf{p} \in \mathcal{X}_{\mathbf{A}} \cap \Delta_{|\mathcal{I}|-1}, \quad (12)$$

where $\Delta_{|\mathcal{I}|-1}$ is the $(|\mathcal{I}| - 1)$ -dimensional simplex. \square

For positive distributions being in $\mathcal{X}_{\mathbf{A}}$ is equivalent to the representations (4), (5), and (6). Therefore, the relational model generated by \mathbf{A} is a subset of the corresponding extended model. For a positive $\boldsymbol{\delta}$, whether or not (4), (5), and (6) hold does not depend on the choice of \mathbf{D} . However, as illustrated next, there exist $\boldsymbol{\delta} \geq \mathbf{0}$, which, due to the pattern of zeros, satisfy (6) for some choice of \mathbf{D} and do not satisfy for another.

Example 2.1 (revisited): The model has dual representations using matrices \mathbf{D}_1 and \mathbf{D}_2 :

$$\mathbf{D}_1 = \begin{pmatrix} 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} \quad \mathbf{D}_2 = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

The distribution $\boldsymbol{\delta} = (0, 0, 0, 1, 1, 1, 0)'$ satisfies (6) if obtained from \mathbf{D}_2 , but does not satisfy (6) if obtained using \mathbf{D}_1 , and therefore, $\boldsymbol{\delta} \notin \mathbf{X}_{\mathbf{A}}$. \square

The support $\text{supp}(\boldsymbol{\delta}) = \{i \in \mathcal{I} : \delta_i > 0\}$ of distributions with zero components which are in $\mathcal{X}_{\mathbf{A}}$ can be characterized using the concept of a facial set which is defined next.

Let $\mathbf{a}_1, \dots, \mathbf{a}_{|\mathcal{I}|}$ denote the columns of \mathbf{A} , and let $C_{\mathbf{A}}$ be the set of all non-negative linear combinations of these columns:

$$C_{\mathbf{A}} = \{\mathbf{t} \in \mathbb{R}_{\geq 0}^J : \exists \boldsymbol{\delta} \in \mathbb{R}_{\geq 0}^{|\mathcal{I}|} \quad \mathbf{t} = \mathbf{A}\boldsymbol{\delta}\}. \quad (13)$$

The relative interior of $C_{\mathbf{A}}$, $\text{relint}(C_{\mathbf{A}})$, comprises such $\mathbf{t} \in \mathbb{R}_{> 0}^J$, for which there exists a (strictly) positive $\boldsymbol{\delta}$ that satisfies $\mathbf{t} = \mathbf{A}\boldsymbol{\delta}$.

The set $C_{\mathbf{A}}$ is a polyhedral cone in \mathbb{R}^J . If an affinely independent set $\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \dots, \mathbf{a}_{i_f}$ of columns of \mathbf{A} spans a proper face of $C_{\mathbf{A}}$, the set of indices $F = \{i_1, i_2, \dots, i_f\}$ is called facial [cf. Grünbaum, 2003, Geiger et al., 2006]. The facial sets of \mathbf{A} are determined by its row space [cf. Fienberg and Rinaldo, 2012]. If $\mathbf{t} \in C_{\mathbf{A}} \setminus \text{relint}(C_{\mathbf{A}})$, then \mathbf{t} is said to lie on a face of $C_{\mathbf{A}}$. In that case, there is a facial set $F = F(\mathbf{t})$, such that

$$\mathbf{t} = s_1 \mathbf{a}_{i_1} + \dots + s_f \mathbf{a}_{i_f}. \quad (14)$$

Equivalently, a set F is facial if and only if there exists a $\mathbf{c} \in \mathbb{R}^J$, such that $\mathbf{c}'\mathbf{a}_i = 0$ for every $i \in F$ and $\mathbf{c}'\mathbf{a}_i > 0$ for every $i \notin F$. The properties of facial sets are formulated in Lemma A.1 given in the Appendix. In particular, only distributions whose support is \mathcal{I} or a facial set of \mathbf{A} may belong to $\mathcal{X}_{\mathbf{A}}$. As an example, the facial sets of the model matrix (9) of AS-independence are $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2, 4\}$, $\{2, 3, 6\}$, $\{1, 3, 5\}$. The support $\{4, 5, 6\}$ of $\boldsymbol{\delta} = (0, 0, 0, 1, 1, 1, 0)'$ from Example 2.1 is not a facial set, and thus $\boldsymbol{\delta}$ cannot be an element of $\mathcal{X}_{\mathbf{A}}$.

The following theorem describes the structure of the parameter set of the extended relational model.

Theorem 3.1. *The extended relational model $\overline{RM}_{\boldsymbol{\delta}}(\mathbf{A})$ is the closure of the relational model $RM_{\boldsymbol{\delta}}(\mathbf{A})$ in the topology of pointwise convergence: $\overline{RM}_{\boldsymbol{\delta}}(\mathbf{A}) = \text{cl}(RM_{\boldsymbol{\delta}}(\mathbf{A}))$.*

The proof is provided in the Appendix. The theorem says that every distribution in the extended model can be obtained as a pointwise limit of a sequence of distributions in the non-extended model. In the following example, such a sequence is found using the construction described in the proof.

Example 2.1 (revisited):

The set $F = \{2, 3, 6\}$ is facial set of \mathbf{A} , and thus, by Lemma A.1, the extended model contains a distribution $\mathbf{p} = (0, p_2, p_3, 0, 0, p_6, 0)'$, where $p_2, p_3, p_6 > 0$ and $p_2 + p_3 + p_6 = 1$. To construct a sequence of distributions in the original model which converges to \mathbf{p} , find θ_2, θ_3 such that

$$\theta_2 = p_2, \quad \theta_3 = p_3, \quad \theta_2 \theta_3 = p_6.$$

From the normalization condition,

$$\theta_2 = \frac{1 - \theta_3}{1 + \theta_3}.$$

Take an arbitrary $\theta_1 \in (0, 1)$, then set

$$\theta_2^{(n)} = \frac{1 - \theta_1 n^{-1} - \theta_3 - \theta_1 n^{-1} \theta_3}{1 + \theta_1 n^{-1} + \theta_3 + \theta_1 n^{-1} \theta_3},$$

and consider

$$\mathbf{p}^{(n)} = (\theta_1 n^{-1}, \theta_2^{(n)}, \theta_3, \theta_1 n^{-1} \theta_2^{(n)}, \theta_1 n^{-1} \theta_3, \theta_2^{(n)} \theta_3, \theta_1 n^{-1} \theta_2^{(n)} \theta_3)'. \quad (14)$$

For every n , $\mathbf{p}^{(n)} \in RM_{\mathbf{p}}(\mathbf{A})$. As $n \rightarrow \infty$, $\theta_2^{(n)} \rightarrow \theta_2$, and therefore, $\mathbf{p}^{(n)} \rightarrow \mathbf{p}$. The construction is complete. \square

An extended relational model can also be defined as a closure of the exponential family corresponding to the original model. The closure of exponential families using the Kullback-Leibler divergence was described for regular families by Brown [1988], among others, and for full families by Csiszár and Matúš [2003]. However, both of these approaches rely on the presence of the overall effect, which implies, through the possibility of normalization, that the Kullback-Leibler divergence is non-negative and Pinsker's inequality [cf. Csiszár, 1975] holds. In the generality considered in the present paper, the approach does not apply, and the Bregman divergence is used to define the closure.

Let $D(\cdot||\cdot)$ denote the Bregman divergence between two vectors $\mathbf{t}, \mathbf{u} \in \mathbb{R}_{\geq 0}^{|\mathcal{I}|}$, associated with the function $f(\mathbf{x}) = \sum_{i \in \mathcal{I}} x(i) \log x(i)$:

$$D(\mathbf{t}||\mathbf{u}) = \sum_{i \in \mathcal{I}} t(i) \log(t(i)/u(i)) + \left(\sum_{i \in \mathcal{I}} u(i) - \sum_{i \in \mathcal{I}} t(i) \right). \quad (15)$$

Under the convention $0 \cdot \log 0 = 0$, $D(\mathbf{t}||\mathbf{u})$ is also defined for non-negative \mathbf{t} and \mathbf{u} if $\text{supp}(\mathbf{t}) \subseteq \text{supp}(\mathbf{u})$. The function $D(\mathbf{t}||\mathbf{u})$ is non-negative, and $D(\mathbf{t}||\mathbf{u}) = 0$ if and only if $\mathbf{t} = \mathbf{u}$. For any $\mathbf{u}^* \in \mathbb{R}_{\geq 0}^{|\mathcal{I}|}$ and for any convex set $\mathcal{S} \subset \mathbb{R}_{\geq 0}^{|\mathcal{I}|}$ there exists a unique $\mathbf{u}^* \in \mathbb{R}_{\geq 0}^{|\mathcal{I}|}$, such that

$$D(\mathbf{u}^*||\mathbf{u}) = \min_{\mathbf{z} \in \mathcal{S}} D(\mathbf{z}||\mathbf{u}), \quad (16)$$

see Bregman [1967]. This \mathbf{u}^* is called the D-projection, or the Bregman projection, of \mathbf{u} on \mathcal{S} . If \mathbf{p}_1 and \mathbf{p}_2 are probability distributions, then $D(\mathbf{p}_1||\mathbf{p}_2)$ is the Kullback-Leibler divergence.

Let $\widetilde{RM}_{\delta}(\mathbf{A})$ be the closure of $RM_{\delta}(\mathbf{A})$ with respect to the Bregman divergence:

$$\widetilde{RM}_{\delta}(\mathbf{A}) = \left\{ \boldsymbol{\delta} \in \bar{\mathcal{P}} : \exists \boldsymbol{\delta}^{(n)} \in RM_{\delta}(\mathbf{A}), n \in \mathbb{N}, \text{ such that } D(\boldsymbol{\delta}||\boldsymbol{\delta}^{(n)}) \rightarrow 0 \text{ as } n \rightarrow \infty \right\}.$$

Theorem 3.2. *The closures of the relational model $RM_{\delta}(\mathbf{A})$ according to the pointwise convergence and to the Bregman divergence coincide.*

Proof. Let $\boldsymbol{\delta}^* \in \widetilde{RM}_{\delta}(\mathbf{A})$. Then, there exists a sequence $\boldsymbol{\delta}^{(n)} \in RM_{\delta}(\mathbf{A})$ such that $\boldsymbol{\delta}^{(n)} \rightarrow \boldsymbol{\delta}^*$ pointwise, as $n \rightarrow \infty$. The function $D(\boldsymbol{\delta}^*||\boldsymbol{\delta}^{(n)})$ is defined and continuous for $\boldsymbol{\delta}^{(n)} > 0$, even if some of the components of $\boldsymbol{\delta}^*$ are zero. Therefore, $D(\boldsymbol{\delta}^*||\boldsymbol{\delta}^{(n)}) \rightarrow 0$, as $n \rightarrow \infty$.

Suppose $\boldsymbol{\delta}^* \in \widetilde{RM}_{\delta}(\mathbf{A})$, and, thus, there exists a sequence $\boldsymbol{\delta}^{(n)} \in RM_{\delta}(\mathbf{A})$, such that:

$$D(\boldsymbol{\delta}^*||\boldsymbol{\delta}^{(n)}) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Therefore, $D(\boldsymbol{\delta}^*||\boldsymbol{\delta}^{(n)}) \leq 1$ for all large enough n . Because the set $\{\boldsymbol{\delta} \geq \mathbf{0} : D(\boldsymbol{\delta}^*||\boldsymbol{\delta}) \leq 1\}$ is compact in $\mathbb{R}^{|\mathcal{I}|}$ [Bregman, 1967], there exists a subsequence $\boldsymbol{\delta}^{(n_k)}$ that converges pointwise to $\boldsymbol{\delta}^*$, as $k \rightarrow \infty$. \square

A relational model $RM_{\delta}(\mathbf{A})$ is a multiplicative family of distributions; the conditions under which the extended model $\overline{RM}_{\delta}(\mathbf{A})$ is also a multiplicative family are studied next.

A distribution $\delta \in \bar{\mathcal{P}}$ is said to factor according to a matrix \mathbf{A} if it has a representation given in (3), with $\theta = (\theta_1, \dots, \theta_J)' \geq \mathbf{0}$. Every distribution in a relational model factors according to the model matrix. However, as the next example demonstrates, an extended model may contain distributions which do not factor according to one choice of the model matrix but do factor according to a different choice.

Example 2.2 (revisited): Any distribution in $RM_{\mathbf{p}}(\mathbf{A})$ factors according to \mathbf{A} , that is,

$$\mathbf{p} = (\theta_1\theta_2\theta_3, \theta_1\theta_2, \theta_1, \theta_3, \theta_1\theta_2\theta_3)', \quad (17)$$

for some $\theta_1, \theta_2, \theta_3 > 0$. The non-negative distribution $\mathbf{p}_0 = (1/8, 1/2, 0, 1/4, 1/8)'$ does not have the multiplicative structure (17), but is in the extended model. To show the latter, take

$$\theta_1^{(n)} = \frac{3}{3n+4}, \quad \theta_2^{(n)} = \frac{n}{2}, \quad \theta_3^{(n)} = \frac{1}{4}, \quad n \geq 1.$$

Then, the sequence

$$\mathbf{p}^{(n)} = \left(\frac{3n}{8(3n+4)}, \frac{3n}{2(3n+4)}, \frac{3}{3n+4}, \frac{1}{4}, \frac{3n}{8(3n+4)} \right)'$$

is in the model, and $\lim_{n \rightarrow \infty} \mathbf{p}^{(n)} = \mathbf{p}_0$. On the other hand, \mathbf{p}_0 factors according to the matrix

$$\mathbf{A}_1 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix},$$

which generates the same extended model as \mathbf{A} does, because $Ker(\mathbf{A}) = Ker(\mathbf{A}_1)$. \square

A necessary and sufficient condition of the existence of such a factorization for a distribution in an extended relational model is given next.

Theorem 3.3. *A distribution $\delta \in \overline{RM}_{\delta}(\mathbf{A})$ factors according to \mathbf{A} if and only if for any $i_0 \notin \text{supp}(\delta)$ there exists an index $j = j(i_0) \in \{1, \dots, J\}$ such that $a_{ji} = 0$ for all $i \in \text{supp}(\delta)$. \square*

The condition of the theorem, called the \mathbf{A} -feasibility of $\text{supp}(\delta)$, means that a generating subset which contains a zero cell of the distribution does not include any positive cell. For extended log-linear models, this condition was proved in Geiger et al. [2006] and Rauh et al. [2011]. The proofs given did not actually rely on the presence of the overall effect and thus apply here.

Maximum likelihood estimation in the extended relational model is studied next.

4 MLE in the extended model

Let F be a facial set, and let \mathbf{A}_F denote the sub-matrix of \mathbf{A} comprising the columns with indices in F , and δ_F denote the sub-vector of δ with indices in F . The following result extends Theorem 9 in Fienberg and Rinaldo [2012].

Theorem 4.1. *Let \mathbf{y} be the vector of observed frequencies under Poisson or multinomial sampling, and let $RM_{\delta}(\mathbf{A})$ be a relational model. Consider \mathbf{q} defined in (7), and assume that $\text{supp}(\mathbf{q}) \subsetneq \mathcal{I}$.*

- (i) *If for all facial sets F , $\text{supp}(\mathbf{q}) \not\subseteq F$, then the MLE $\tilde{\delta}_{\mathbf{y}}$ under the model $\overline{RM}_{\delta}(\mathbf{A})$ exists, and is also the MLE under $RM_{\delta}(\mathbf{A})$: $\tilde{\delta}_{\mathbf{y}} = \hat{\delta}_{\mathbf{y}}$. Otherwise,*
- (ii) *Let F be the smallest facial set such that $\text{supp}(\mathbf{q}) \subseteq F$. Then the MLE $\hat{\delta}_{\mathbf{y},F}$ of δ_F under the model $RM_{\delta_F}(\mathbf{A}_F)$ exists, and $\tilde{\delta}_{\mathbf{y}} = (\hat{\delta}_{\mathbf{y},F}, \mathbf{0}_{\mathcal{I} \setminus F})$ is the MLE under the model $\overline{RM}_{\delta}(\mathbf{A})$.*
- (iii) *The MLE $\tilde{\delta}_{\mathbf{y}}$ under $\overline{RM}_{\delta}(\mathbf{A})$ always exists and is the unique point of $\mathcal{X}_{\mathbf{A}}$ which satisfies:*

$$\begin{aligned} \mathbf{A}\delta &= \gamma \mathbf{A}\mathbf{q}, \text{ for some } \gamma > 0; \\ \mathbf{1}'\delta &= 1 \quad (\text{only for } \delta \equiv \mathbf{p}). \end{aligned} \tag{18}$$

The vector $\tilde{\delta}_{\mathbf{y}}$ is called the extended MLE of δ under the relational model. The proof is given in the Appendix. The following example illustrates the theorem.

Example 2.2 (revisited):

Notice first that $F = \{1, 2, 4, 5\}$ is a facial set of \mathbf{A} . The support of the observed distribution $\text{supp}(\mathbf{q}) = \{1, 2, 4\}$ is a subset of F . Therefore, the MLE of \mathbf{q} exists in the closure of the relational model. As it was shown earlier, the distribution $\mathbf{p}_0 = (1/8, 1/2, 0, 1/4, 1/8)'$ is in $\overline{RM}_{\mathbf{p}}(\mathbf{A})$. As $\mathbf{A}\mathbf{p}_0 = 7/8\mathbf{A}\mathbf{q}$, the extended MLE of \mathbf{q} is \mathbf{p}_0 . \square

The next theorem establishes a condition under which the maximum likelihood estimates of the model parameters under an extended relational model exist:

Theorem 4.2. *Assume that the MLE $\hat{\delta}$ under the extended relational model $\overline{RM}_{\delta}(\mathbf{A})$ exists. The maximum likelihood estimates of the model parameters θ exist if and only if $\text{supp}(\hat{\delta})$ is \mathbf{A} -feasible.*

Proof. By Theorem 3.3, the distribution $\hat{\delta}$ factors according to \mathbf{A} if and only if $\text{supp}(\hat{\delta})$ is \mathbf{A} -feasible. In this case $\hat{\delta}(i) = \prod_{j=1}^J \hat{\theta}_j^{a_{ij}}$ for all $i \in \mathcal{I}$, and, by uniqueness, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_J)'$ are the maximum likelihood estimates of the model parameters. \square

If $\text{supp}(\hat{\delta})$ is not \mathbf{A} -feasible, then $\hat{\delta}$ is the limit of a sequence of the positive distributions in the model which factor according to \mathbf{A} . Although the cell parameters of these distributions can be factored using some model parameters $\theta^{(n)} > \mathbf{0}$, the limits of individual components of $\theta^{(n)}$, as $n \rightarrow \infty$, may not exist. In the case of the log-linear models this fact was illustrated by Rinaldo [2006]. The same situation occurs in the construction of Example 2.2, where $\theta_2^{(n)} \rightarrow \infty$ as $n \rightarrow \infty$.

As Theorem 4.1 implies, the MLE in the extended relational model can be obtained using the MLE in a non-extended model. Klimova and Rudas [2015] proposed a generalized iterative scaling procedure, called G-IPF, for computing the MLE under (non-extended) relational models. The algorithm relies on the condition that $\mathbf{A}\mathbf{q} > \mathbf{0}$. Every iteration of this procedure implements the following algorithm, IPF(γ), for a specific value of γ .

IPF(γ) Algorithm:

Set $n = 0$; $\delta_\gamma^{(0)}(i) = 1$ for all $i \in \mathcal{I}$, and proceed as follows.

Step 1: Find $j \in \{1, 2, \dots, J\}$, such that $n + 1 \equiv j \pmod{J}$;

Step 2: Compute

$$\delta_\gamma^{(n+1)}(i) = \delta_\gamma^{(n)}(i) \left(\gamma \frac{A_j \mathbf{q}}{A_j \delta_\gamma^{(n)}} \right)^{a_{ji}} \quad \text{for all } i \in \mathcal{I}. \quad (19)$$

Step 3: While $\gamma A_j \mathbf{q} \neq A_j \delta_\gamma^{(n+1)}$ for at least one j , set $n = n + 1$, go to **Step 1**.

Step 4: Set $\delta_\gamma^* = \delta_\gamma^{(n)}$, and finish. \square

The G-IPF algorithm commences with executing IPF(γ) for $\gamma = 1$, which is sufficient to compute the MLE in the case of probabilities with the overall effect and in the case of intensities. If in the case of probabilities the overall effect is not present, G-IPF updates γ and calls IPF(γ) again. The procedure is repeated until, for some γ , the limit vector δ_γ^* sums to 1, and thus is a parameter of a non-negative probability distribution. The variant of G-IPF, which employs the bisection method to update γ , is described in the following.

G-IPF Algorithm:

If $\delta \equiv \lambda$, compute $\tilde{\lambda}$ using IPF(1), and finish.

If $\delta \equiv \mathbf{p}$, compute \mathbf{p}^* using IPF(1).

If $\mathbf{1p}^* = 1$, set $\tilde{\mathbf{p}} = \mathbf{p}^*$, and finish. Otherwise, compute $\gamma_L = (\mathbf{1}' \mathbf{A} \mathbf{q})^{-1}$, $\gamma_R = \min \{1/A_1 \mathbf{q}, \dots, 1/A_J \mathbf{q}\}$, and proceed as follows:

Step 1: Find $\delta_{(\gamma_L + \gamma_R)/2}^*$ using IPF(γ).

Step 2: While $\mathbf{1} \delta_{(\gamma_L + \gamma_R)/2}^* \neq 1$,

if $\mathbf{1} \delta_{(\gamma_L + \gamma_R)/2}^* < 1$, set $\gamma_L = \frac{\gamma_L + \gamma_R}{2}$,

else set $\gamma_R = \frac{\gamma_L + \gamma_R}{2}$;

go to **Step 1**.

Step 3: Set $\tilde{\mathbf{p}} = \delta_{(\gamma_L + \gamma_R)/2}^*$, and finish. \square

If $\mathbf{A} \mathbf{q} > \mathbf{0}$, the G-IPF algorithm applies to the extended case directly.

Theorem 4.3. *Let \mathbf{y} be the vector of observed frequencies under Poisson or multinomial sampling, with \mathbf{q} defined in (7), and let $RM_\delta(\mathbf{A})$ be a relational model. Assume that $\mathbf{A} \mathbf{q} > \mathbf{0}$. The G-IPF algorithm converges to the MLE $\tilde{\delta}_\mathbf{y}$ under $\overline{RM}_\delta(\mathbf{A})$.*

Proof. As $\mathbf{A}\mathbf{q} > \mathbf{0}$, the IPF-sequence $\boldsymbol{\delta}_\gamma^{(n)}$ defined in (19) is positive, and the proof of its convergence in Klimova and Rudas [2015, Theorem 3.2] applies. In particular, the limit of the sequence, $\boldsymbol{\delta}_\gamma^*$, satisfies $\mathbf{A}\boldsymbol{\delta}_\gamma^* = \gamma\mathbf{A}\mathbf{q}$, and, for an arbitrary kernel basis matrix \mathbf{D} , $\mathbf{D}\log \boldsymbol{\delta}_\gamma^{(n)} = \mathbf{0}$ for all $n \in \mathbb{Z}_{\geq 0}$. The latter implies that $\boldsymbol{\delta}_\gamma^{(n)} \in \mathcal{X}_\mathbf{A}$ for all n , and, as $\mathcal{X}_\mathbf{A}$ is a closed set in $\mathbb{R}_{\geq 0}^{|\mathcal{I}|}$, $\boldsymbol{\delta}_\gamma^* \in \mathcal{X}_\mathbf{A}$.

Let $\boldsymbol{\delta}_1^*$ be the limit vector obtained from IPF(1), and thus $\boldsymbol{\delta}_1^* \in \mathcal{X}_\mathbf{A}$ and $\mathbf{A}\boldsymbol{\delta}_1^* = \mathbf{A}\mathbf{q}$.

Suppose $\boldsymbol{\delta} \equiv \boldsymbol{\lambda}$. Then, as (18) holds for $\boldsymbol{\delta}_1^*$ with $\gamma = 1$, Theorem 4.1(iii) implies that $\boldsymbol{\delta}_1^*$ is equal to the extended MLE: $\tilde{\boldsymbol{\delta}}_\mathbf{y} = \boldsymbol{\delta}_1^*$.

Suppose $\boldsymbol{\delta} \equiv \mathbf{p}$. First, assume that the overall effect is present, and thus there exists a $\mathbf{k} \in \mathbb{R}_{\geq 0}^J$, such that $\mathbf{1}' = \mathbf{k}'\mathbf{A}$. The latter yields that $\mathbf{1}'\boldsymbol{\delta}_1^* = \mathbf{k}'\mathbf{A}\boldsymbol{\delta}_1^* = \mathbf{k}'\mathbf{A}\mathbf{q} = \mathbf{1}'\mathbf{q} = 1$. Therefore, (18) holds for $\boldsymbol{\delta}_1^*$ with $\gamma = 1$. By Theorem 4.1(iii), $\tilde{\boldsymbol{\delta}}_\mathbf{y} = \boldsymbol{\delta}_1^*$.

Now, assume that the overall effect is not present. In this situation, G-IPF updates γ and calls IPF(γ); and this procedure is repeated until a γ^* for which the IPF-limit $\boldsymbol{\delta}_{\gamma^*}^*$ sums to 1 is found. Then, $\boldsymbol{\delta}_{\gamma^*}^*$ satisfies (18) with $\gamma = \gamma^*$. By Theorem 4.1(iii), $\tilde{\boldsymbol{\delta}}_\mathbf{y} = \boldsymbol{\delta}_{\gamma^*}^*$. \square

Next, it is shown how G-IPF can be used if the condition $\mathbf{A}\mathbf{q} > \mathbf{0}$ does not hold. Let $\mathcal{J}_0 = \{j \in \{1, \dots, J\} : A_j\mathbf{q} = \mathbf{0}\}$, and assume that $\mathcal{J}_0 \neq \emptyset$. Further, let $\mathcal{I}_0 = \{i \in \mathcal{I} : \exists j \in \mathcal{J}_0 \text{ } a_{ji} = 1\}$, and let $\mathcal{I}_* = \mathcal{I} \setminus \mathcal{I}_0$. Denote by \mathbf{A}_* the matrix obtained from \mathbf{A} by removing the columns with indices in \mathcal{I}_0 and by removing the zero rows, if such occur afterwards, and by $\boldsymbol{\delta}_*$, \mathbf{y}_* , and \mathbf{q}_* the corresponding sub-vectors of $\boldsymbol{\delta}$, \mathbf{y} , and \mathbf{q} . By Theorem 4.1(iii), the MLE $\tilde{\boldsymbol{\delta}}_{\mathbf{y}_*}$ of \mathbf{y}_* under $\overline{RM}_{\boldsymbol{\delta}_*}(\mathbf{A}_*)$ exists and is unique. Since $\mathbf{A}_*\mathbf{q}_* > \mathbf{0}$, $\tilde{\boldsymbol{\delta}}_{\mathbf{y}_*}$ can be computed using G-IPF, see Theorem 4.3, and the following holds:

Theorem 4.4. *The MLE of \mathbf{y} under $\overline{RM}_{\boldsymbol{\delta}}(\mathbf{A})$ is equal to $\tilde{\boldsymbol{\delta}}_\mathbf{y} = (\tilde{\boldsymbol{\delta}}_{\mathbf{y}_*}, \mathbf{0}_{\mathcal{I}_0})$.*

Proof. In order to show that $\tilde{\boldsymbol{\delta}}_\mathbf{y} \in \mathcal{X}_\mathbf{A}$, it will first be verified that \mathcal{I}_* is a facial set of \mathbf{A} . Let \mathbf{a}_i be the i -th column of \mathbf{A} , then, with $\mathbf{c} = (\mathbf{0}_{\mathcal{J} \setminus \mathcal{J}_0}, \mathbf{1}_{\mathcal{J}_0})'$, $\mathbf{c}'\mathbf{a}_i = 0$ for any $i \in \mathcal{I}_*$. If $i \notin \mathcal{I}_*$, then $a_{ji} = 1$ for some $j \in \mathcal{J}_0$, and thus $\mathbf{c}'\mathbf{a}_i > 0$. Therefore, \mathcal{I}_* is a facial set of \mathbf{A} . Then, by Lemma A.3, $\tilde{\boldsymbol{\delta}}_\mathbf{y} \in \mathcal{X}_\mathbf{A}$.

Next, in the case of probabilities, the normalization condition $\mathbf{1}'_{\mathcal{I}_*} \tilde{\boldsymbol{\delta}}_{\mathbf{y}_*} = 1$ implies that $\mathbf{1}'\tilde{\boldsymbol{\delta}}_\mathbf{y} = 1$. Further, $\mathbf{A}_*\tilde{\boldsymbol{\delta}}_{\mathbf{y}_*} = \gamma\mathbf{A}_*\mathbf{q}_*$ implies that $\mathbf{A}\tilde{\boldsymbol{\delta}}_\mathbf{y} = \gamma\mathbf{A}\mathbf{q}$.

Finally, by Theorem 4.1(iii), $\tilde{\boldsymbol{\delta}}_\mathbf{y}$ is the MLE of \mathbf{y} under $\overline{RM}_{\boldsymbol{\delta}}(\mathbf{A})$. \square

5 Conclusion

Some research areas deal with populations of a complex structure to which inference based on the standard log-linear approach does not apply, but the relational model framework can be used. The relational models are more flexible as they allow effects associated with arbitrary subsets of cells, can be used for incomplete tables, and do not require the presence of an overall effect. Similarly to the log-linear case, data with zero counts may not possess an MLE under a relational model. A necessary and sufficient condition for the existence of the MLE was obtained in Section 2. When this condition does not hold, an MLE may exist in the extended sense, that is, in the closure of the relational model. Different but equivalent ways of defining such a closure, and a necessary and sufficient condition for the existence of

the extended MLE in it were presented in Section 3. A condition under which a distribution in the closure factorizes according to the model matrix was also given. These results were obtained using concepts and methods of algebraic statistics. Just like in the case of relational models, the cases of multinomial and Poisson sampling are not equivalent. It was shown in Section 4, that the generalized relative proportional fitting procedure originally suggested for relational models also works when the data contain zeros and the MLE is sought for in the closure of a relational model.

A Appendix

A.1 Properties of facial sets

Lemma A.1. *Let \mathbf{A} be the model matrix of a relational model, and let F be a facial set of \mathbf{A} . Then:*

- (i) *There exists a $\mathbf{c} \in \mathbb{R}^J$, such that $\mathbf{c}'\mathbf{a}_i = 0$ for any $i \in F$ and $\mathbf{c}'\mathbf{a}_i > 0$ for any $i \notin F$.*
- (ii) *For any $\mathbf{d} \in \text{Ker}(\mathbf{A})$, either both $\text{supp}(\mathbf{d}^+) \subseteq F$ and $\text{supp}(\mathbf{d}^-) \subseteq F$ or both $\text{supp}(\mathbf{d}^+) \not\subseteq F$ and $\text{supp}(\mathbf{d}^-) \not\subseteq F$.*
- (iii) *For any $\boldsymbol{\delta} \in \mathcal{X}_{\mathbf{A}}$, either $\text{supp}(\boldsymbol{\delta}) = \mathcal{I}$ or $\text{supp}(\boldsymbol{\delta})$ is a facial set of \mathbf{A} .*
- (iv) *If F is a facial set of \mathbf{A} , there exists a $\boldsymbol{\delta} \in \mathcal{X}_{\mathbf{A}}$, such that $\text{supp}(\boldsymbol{\delta}) = F$.*

The statements of the lemma were proved by Geiger et al. [2006] and Rauh, Kahle, and Ay [2011] for models of type (2) when the overall effect is present. Their proofs do not rely on the latter characteristic and thus apply here.

The next lemma shows that the condition of existence of the MLE given in Theorem 2.2 can also be formulated in terms of facial sets.

Lemma A.2. *There exists a $\mathbf{z} > \mathbf{0}$, such that $\mathbf{Az} = \mathbf{Aq}$, if and only if $\text{supp}(\mathbf{q})$ is not contained in any facial set of \mathbf{A} .*

Proof. Suppose there exists a $\mathbf{z} > \mathbf{0}$, such that $\mathbf{Az} = \mathbf{Aq}$, and thus $\mathbf{d} = \mathbf{z} - \mathbf{q} \in \text{Ker}(\mathbf{A})$ and $\mathbf{q} + \mathbf{d} > \mathbf{0}$.

Let F be a facial set of \mathbf{A} . If both $\mathbf{d}^+ \subseteq F$ and $\mathbf{d}^- \subseteq F$, then $d_i = 0$ for all $i \notin F$. Because $\mathbf{q} + \mathbf{d} > \mathbf{0}$, $q_i + d_i = q_i > 0$ for all $i \notin F$. Therefore, $\text{supp}(\mathbf{q})$ is not contained in F . Otherwise, see Lemma A.1, both $\mathbf{d}^+ \not\subseteq F$ and $\mathbf{d}^- \not\subseteq F$, and there exists an $i \notin F$ such that $d_i < 0$. If q_i was zero, then $q_i + d_i$ would be negative, which contradicts the initial assumption $\mathbf{q} + \mathbf{d} > \mathbf{0}$. Therefore, q_i has to be positive, which implies that $\text{supp}(\mathbf{q})$ is not contained in F .

To prove the converse, assume that $\text{supp}(\mathbf{q})$ is not contained in any facial set F . Suppose the equation $\mathbf{Aq} = \mathbf{Az}$ has no (strictly) positive solution in \mathbf{z} , and, therefore, $\mathbf{Aq} \notin \text{relint}(C_{\mathbf{A}})$. A non-negative solution always exists, and thus \mathbf{Aq} belongs to a face of $C_{\mathbf{A}}$. Then (14) holds for $\mathbf{t} = \mathbf{Aq}$ for some facial set F ; without loss of generality, $F = \{1, \dots, f\}$:

$$\mathbf{Aq} = s_1 \mathbf{a}_1 + \dots + s_f \mathbf{a}_f.$$

Hence,

$$(q_1 - s_1)\mathbf{a}_1 + \cdots + (q_f - s_f)\mathbf{a}_f + q_{f+1}\mathbf{a}_{f+1} + \cdots + q_{|\mathcal{I}|}\mathbf{a}_{|\mathcal{I}|} = \mathbf{0}. \quad (20)$$

Multiplying both sides of (20) by a vector \mathbf{c} , such that $\mathbf{c}'\mathbf{a}_i = 0$ for $i \in F$ and $\mathbf{c}'\mathbf{a}_i > 0$ for $i \notin F$, leads to:

$$q_{f+1} = 0, \dots, q_{|\mathcal{I}|} = 0,$$

which means that $\text{supp}(\mathbf{q}) \subset F$. This contradicts the initial assumption that $\text{supp}(\mathbf{q})$ is not contained in any facial set. \square

The following lemma is used in the proofs of Theorems 4.1 and 4.4.

Lemma A.3. *If F is a facial set of \mathbf{A} , then, for any $\boldsymbol{\delta}_F \in \mathcal{X}_{\mathbf{A}_F}$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_F, \mathbf{0}_{\mathcal{I} \setminus F}) \in \mathcal{X}_{\mathbf{A}}$.*

Proof. Take an arbitrary $\mathbf{d} \in \text{Ker}(\mathbf{A})$. As F is a facial set of \mathbf{A} , by Lemma A.1(ii), exactly one of the following holds:

$$\text{supp}(\mathbf{d}^+) \subseteq F \text{ and } \text{supp}(\mathbf{d}^-) \subseteq F, \text{ or } \text{supp}(\mathbf{d}^+) \not\subseteq F \text{ and } \text{supp}(\mathbf{d}^-) \not\subseteq F.$$

In the first case, there exists a $\mathbf{d}_F \in \text{Ker}(\mathbf{A}_F)$, such that $\mathbf{d} = (\mathbf{d}_F, \mathbf{0}_{\mathcal{I} \setminus F})$. Since $\boldsymbol{\delta}_F \in \mathcal{X}_{\mathbf{A}_F}$, $(\boldsymbol{\delta}_F)^{\mathbf{d}_F^+} = (\boldsymbol{\delta}_F)^{\mathbf{d}_F^-}$, and, therefore,

$$(\boldsymbol{\delta})^{\mathbf{d}^+} = (\boldsymbol{\delta}_F)^{\mathbf{d}_F^+} \cdot (\mathbf{0}_{\mathcal{I} \setminus F})^{\mathbf{0}_{\mathcal{I} \setminus F}} = (\boldsymbol{\delta}_F)^{\mathbf{d}_F^-} \cdot (\mathbf{0}_{\mathcal{I} \setminus F})^{\mathbf{0}_{\mathcal{I} \setminus F}} = (\boldsymbol{\delta})^{\mathbf{d}^-}.$$

In the second case, there exist such $i_1, i_2 \notin F$ that $d_{i_1} > 0$ and $d_{i_2} < 0$, and thus,

$$(\boldsymbol{\delta})^{\mathbf{d}^+} = (\boldsymbol{\delta}_F)^{\mathbf{d}_F^+} \cdot 0 = (\boldsymbol{\delta}_F)^{\mathbf{d}_F^-} \cdot 0 = (\boldsymbol{\delta})^{\mathbf{d}^-}.$$

As $(\boldsymbol{\delta})^{\mathbf{d}^+} = (\boldsymbol{\delta})^{\mathbf{d}^-}$ for any $\mathbf{d} \in \text{Ker}(\mathbf{A})$, $\boldsymbol{\delta} \in \mathcal{X}_{\mathbf{A}}$. \square

A.2 Proof of Theorem 3.1

The proof extends the arguments given by Geiger et al. [2006] and Rauh et al. [2011]. It will be shown first that for any distribution in $\overline{RM}_\delta(\mathbf{A})$ there exists a sequence of distributions in $RM_\delta(\mathbf{A})$ that converges to it pointwise.

Let $\boldsymbol{\delta}^* \in \overline{RM}_\delta(\mathbf{A})$. By Lemma A.1, as $\boldsymbol{\delta}^* \in \mathcal{X}_{\mathbf{A}}$, $F = \text{supp}(\boldsymbol{\delta}^*)$ is either \mathcal{I} or a facial set of \mathbf{A} . If $F = \mathcal{I}$, then $\boldsymbol{\delta}^* > \mathbf{0}$, and the statement holds with $\boldsymbol{\delta}^{(n)} \equiv \boldsymbol{\delta}^*$. Assume that $F \subsetneq \mathcal{I}$. For simplicity of exposition, let $F = \{1, \dots, f\}$, and then $\boldsymbol{\delta}^* = (\delta_1^*, \dots, \delta_f^*, 0, \dots, 0)$.

First, find $\eta_1, \dots, \eta_J > 0$ that satisfy:

$$\prod_{j=1}^J \eta_j^{a_{ji}} = \delta_i^* \quad \text{for } i \in F.$$

The existence of such θ 's can be proved using the same argument as Geiger et al. [2006, p.28] gave for the case of extended log-linear models. By Lemma A.1, there exists a $\mathbf{c} = (c_1, \dots, c_J)' \in \mathbb{R}^J$, such that $\mathbf{c}'\mathbf{a}_i = 0$ for all $i \in F$ and $\mathbf{c}'\mathbf{a}_i > 0$ for any $i \notin F$. Order the columns of \mathbf{A} so that $c_1 > 0$, and then order the rows of \mathbf{A} so that $a_{11} = 1$.

If $\boldsymbol{\delta} \equiv \boldsymbol{\lambda}$, set, for $n \in \mathbb{Z}_{>0}$,

$$\lambda_i^{(n)} = \prod_{j=1}^J (n^{-c_j} \eta_j)^{a_{ji}}, \quad i \in \mathcal{I}.$$

The distribution $\boldsymbol{\lambda}^{(n)} = (\lambda_1^{(n)}, \dots, \lambda_{|\mathcal{I}|}^{(n)})'$ is positive and satisfies (3) with $\theta_j = n^{-c_j} \eta_j$. Therefore, $\boldsymbol{\lambda}^{(n)} \in RM_{\boldsymbol{\lambda}}(\mathbf{A})$. Further,

$$\lim_{n \rightarrow \infty} \lambda_i^{(n)} = \lim_{n \rightarrow \infty} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=1}^J \eta_j^{a_{ji}} = \begin{cases} \delta_i^*, & \text{if } i \in F, \\ 0, & \text{if } i \notin F, \end{cases}$$

thus $\boldsymbol{\lambda}^{(n)} \rightarrow \boldsymbol{\delta}^*$ pointwise, as $n \rightarrow \infty$.

If $\boldsymbol{\delta} \equiv \mathbf{p}$, take

$$\eta_1^{(n)} = \frac{1 - \sum_{i: a_{1i}=0} \prod_{j=2}^J (n^{-c_j} \eta_j)^{a_{ji}}}{\sum_{i: a_{1i}=1} \prod_{j=2}^J (n^{-c_j} \eta_j)^{a_{ji}}},$$

and set

$$p_i^{(n)} = (\eta_1^{(n)})^{a_{1i}} \prod_{j=2}^J (n^{-c_j} \eta_j)^{a_{ji}}, \quad i \in \mathcal{I}.$$

The choice of $\eta_1^{(n)}$ implies that $\mathbf{1}'\mathbf{p}^{(n)} = 1$. As $\mathbf{p}^{(n)} = (p_1^{(n)}, \dots, p_{|\mathcal{I}|}^{(n)})'$ is positive and satisfies (3) with $\theta_1 = \eta_1^{(n)}$, $\theta_j = n^{-c_j} \eta_j$, for $j = 2, \dots, J$, $\mathbf{p}^{(n)} \in RM_{\mathbf{p}}(\mathbf{A})$. Next, because $\mathbf{c}'\mathbf{a}_i = 0$ if $i \in F$,

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{c_1} \eta_1^{(n)} &= \lim_{n \rightarrow \infty} \frac{n^{c_1} (1 - \sum_{a_{1i}=0, i \in F} \prod_{j=2}^J \eta_j^{a_{ji}} - \sum_{a_{1i}=0, i \notin F} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \eta_j^{a_{ji}})}{n^{c_1} (\sum_{a_{1i}=1, i \in F} \prod_{j=2}^J \eta_j^{a_{ji}} + \sum_{a_{1i}=1, i \notin F} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \eta_j^{a_{ji}})} \\ &= \frac{1 - \sum_{i \in F: a_{1i}=0} \prod_{j=2}^J \eta_j^{a_{ji}}}{\sum_{i \in F: a_{1i}=1} \prod_{j=2}^J \eta_j^{a_{ji}}} = \eta_1. \end{aligned} \quad (21)$$

Further, for $i \in \mathcal{I}$, using (21),

$$\begin{aligned} \lim_{n \rightarrow \infty} p_i^{(n)} &= \lim_{n \rightarrow \infty} n^{a_{1i}c_1 - \mathbf{c}'\mathbf{a}_i} (\eta_1^{(n)})^{a_{1i}} \prod_{j=2}^J \eta_j^{a_{ji}} = \lim_{n \rightarrow \infty} n^{-\mathbf{c}'\mathbf{a}_i} (n^{c_1} \eta_1^{(n)})^{a_{1i}} \prod_{j=2}^J \eta_j^{a_{ji}} \\ &= \lim_{n \rightarrow \infty} n^{-\mathbf{c}'\mathbf{a}_i} (\eta_1)^{a_{1i}} \prod_{j=2}^J \eta_j^{a_{ji}} = \lim_{n \rightarrow \infty} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=1}^J \eta_j^{a_{ji}} = \begin{cases} \delta_i^* & i \in F, \\ 0 & i \notin F. \end{cases} \end{aligned}$$

Hence, $\mathbf{p}^{(n)} \rightarrow \underline{\boldsymbol{\delta}^*}$ pointwise, as $n \rightarrow \infty$.

Therefore, $RM_{\boldsymbol{\delta}}(\mathbf{A}) \subset cl(RM_{\boldsymbol{\delta}}(\mathbf{A}))$.

To prove the converse, choose a $\boldsymbol{\delta}^* \in cl(RM_{\boldsymbol{\delta}}(\mathbf{A}))$. Then, $\boldsymbol{\delta}^*$ is a pointwise limit of a sequence of distributions in $RM_{\boldsymbol{\delta}}(\mathbf{A})$, and $\boldsymbol{\delta}^*$ is the pointwise limit of a sequence in $\mathcal{X}_{\mathbf{A}}$. As $\mathcal{X}_{\mathbf{A}}$ is closed in the topology of pointwise convergence [cf. Geiger et al., 2006], $\boldsymbol{\delta}^* \in \mathcal{X}_{\mathbf{A}}$. If $\boldsymbol{\delta} \equiv \mathbf{p}$, both $\boldsymbol{\delta}^*$ and the sequence converging to it belong to the simplex $\Delta_{|\mathcal{I}|-1}$. Therefore, $\boldsymbol{\delta}^* \in \overline{RM_{\boldsymbol{\delta}}(\mathbf{A})}$, and the proof is complete. \square

A.3 Proof of Theorem 4.1:

The statement (i) follows from Theorem 2.2 and Lemma A.2. \square

In order to prove (ii), notice first that, the smallest facial set F of \mathbf{A} which contains $\text{supp}(\mathbf{q})$ is uniquely defined. In this case, $\mathbf{A}_F \mathbf{q}_F \in \text{relint}(\mathbf{C}_{\mathbf{A}_F})$, and, therefore, $\text{supp}(\mathbf{q})$ is not contained in any facial set of \mathbf{A}_F . By part (i) of this theorem, the MLE $\hat{\delta}_{\mathbf{y}_F}$ under $RM_{\delta_F}(\mathbf{A}_F)$ exists.

Let $\tilde{\delta}_{\mathbf{y}} = (\hat{\delta}_{\mathbf{y}_F}, \mathbf{0}_{\mathcal{I} \setminus F})$. By Lemma A.3, $\tilde{\delta}_{\mathbf{y}} \in \mathcal{X}_{\mathbf{A}}$. If $\delta \equiv \mathbf{p}$, $\mathbf{1}' \hat{\mathbf{p}}_{\mathbf{y}_F} = 1$, and thus $\tilde{\mathbf{p}}_{\mathbf{y}}$ satisfies the normalization condition $\mathbf{1}' \tilde{\mathbf{p}}_{\mathbf{y}} = 1$. It will be shown next that $\tilde{\delta}_{\mathbf{y}}$ maximizes the full log-likelihood of \mathbf{y} .

Let $\delta \equiv \lambda$. The log-likelihood under the model $RM_{\lambda_F}(\mathbf{A}_F)$ is equal to

$$l_F(\mathbf{q}_F, \lambda_F) = \sum_{i \in F} q_{Fi} \log \lambda_{Fi} - \sum_{i \in F} \lambda_{Fi},$$

and for any $\lambda_F > 0$, $l_F(\mathbf{q}_F, \lambda_F) \leq l_F(\mathbf{q}_F, \hat{\lambda}_{\mathbf{y}_F})$.

Let $\lambda = (\lambda'_F, \mathbf{0})'$, and let $\lambda^{(n)}$ be the sequence that was described in the proof of Theorem 3.1. The full log-likelihood of the elements of this sequence is

$$\begin{aligned} l(\mathbf{q}, \lambda^{(n)}) &= \sum_{i \in \mathcal{I}} q_i \log \lambda_i^{(n)} - \sum_{i \in \mathcal{I}} \lambda_i^{(n)} = \sum_{i \in F} q_i \log \lambda_i^{(n)} - \sum_{i \in \mathcal{I}} \lambda_i^{(n)} \\ &= \sum_{i \in F} q_i \log \{n^{-c\mathbf{a}_i} \prod_{j=1}^J \theta_j^{a_{ji}}\} - \sum_{i \in \mathcal{I}} n^{-c\mathbf{a}_i} \prod_{j=1}^J \theta_j^{a_{ji}} \\ &= \sum_{i \in F} q_i \log \{\prod_{j=1}^J \theta_j^{a_{ji}}\} - \sum_{i \in F} \prod_{j=1}^J \theta_j^{a_{ji}} - \sum_{i \notin F} n^{-c\mathbf{a}_i} \prod_{j=1}^J \theta_j^{a_{ji}} \\ &= l_F(\mathbf{q}_F, \lambda_F) - \sum_{i \notin F} n^{-c\mathbf{a}_i} \prod_{j=1}^J \theta_j^{a_{ji}}. \end{aligned}$$

Therefore,

$$l(\mathbf{q}, \lambda^{(n)}) \leq l_F(\mathbf{q}_F, \lambda_F) \leq l_F(\mathbf{q}_F, \hat{\lambda}_{\mathbf{y}_F}). \quad (22)$$

Let $\delta \equiv \mathbf{p}$. The log-likelihood under the model $RM_{\mathbf{p}_F}(\mathbf{A}_F)$ is equal to

$$l_F(\mathbf{q}_F, \mathbf{p}_F) = \sum_{i=1}^f q_{Fi} \log p_{Fi},$$

and for any $\mathbf{p}_F > 0$, such that $\mathbf{1}' \mathbf{p}_F = 1$, $l_F(\mathbf{q}_F, \mathbf{p}_F) \leq l_F(\mathbf{q}_F, \hat{\mathbf{p}}_{\mathbf{y}_F})$.

Let $\mathbf{p} = (\mathbf{p}'_F, \mathbf{0})'$, and let $\mathbf{p}^{(n)}$ be the sequence that was described in the proof of Theorem

3.1. The full log-likelihood of the elements of this sequence is

$$\begin{aligned}
l(\mathbf{q}, \mathbf{p}^{(n)}) &= \sum_{i \in \mathcal{I}} q_i \log p_i^{(n)} = \sum_{i \in F} q_i \log p_i^{(n)} \\
&= \sum_{i \in F} q_i \log \{(\theta_1^{(n)})^{a_{1i}} \prod_{j=2}^J (n^{-c_j} \theta_j^{a_{ji}})\} = \sum_{i \in F} q_i \log \{(\theta_1^{(n)})^{a_{1i}} n^{a_{1i}c_1 - \mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \theta_j^{a_{ji}}\} \\
&= \sum_{i \in F: a_{i1}=1} q_i \log \theta_1^{(n)} n^{c_1} \prod_{j=2}^J \theta_j^{a_{ji}} + \sum_{i \in F: a_{i1}=0} q_i \log \prod_{j=2}^J \theta_j^{a_{ji}} \\
&= \sum_{i \in F: a_{i1}=1} q_i \log \prod_{j=1}^J \theta_j^{a_{ji}} + \sum_{i \in F: a_{i1}=0} q_i \log \prod_{j=1}^J \theta_j^{a_{ji}} - \sum_{i \in F: a_{i1}=1} q_i \log \{\theta_1 / (\theta_1^{(n)} n^{c_1})\} \\
&= l_F(\mathbf{q}_F, \mathbf{p}_F) - \log \{\theta_1 / (\theta_1^{(n)} n^{c_1})\} \cdot \sum_{i \in F: a_{i1}=1} q_i.
\end{aligned}$$

It will be shown next that $\theta_1 / (\theta_1^{(n)} n^{c_1}) > 1$.

$$\begin{aligned}
\frac{\theta_1}{\theta_1^{(n)} n^{c_1}} &= \frac{1 - \sum_{i \in F: a_{i1}=0} \prod_{j=2}^J \theta_j^{a_{ji}}}{\sum_{i \in F: a_{i1}=1} \prod_{j=2}^J \theta_j^{a_{ji}}} \\
&\cdot \frac{n^{c_1} (\sum_{a_{i1}=1, i \in F} \prod_{j=2}^J \theta_j^{a_{ji}} + \sum_{a_{i1}=1, i \notin F} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \theta_j^{a_{ji}})}{n^{c_1} (1 - \sum_{a_{i1}=0, i \in F} \prod_{j=2}^J \theta_j^{a_{ji}} - \sum_{a_{i1}=0, i \notin F} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \theta_j^{a_{ji}})} \\
&= \left(1 + \frac{\sum_{a_{i1}=1, i \notin F} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \theta_j^{a_{ji}}}{\sum_{i \in F: a_{i1}=1} \prod_{j=2}^J \theta_j^{a_{ji}}} \right) / \left(1 - \frac{\sum_{a_{i1}=0, i \notin F} n^{-\mathbf{c}'\mathbf{a}_i} \prod_{j=2}^J \theta_j^{a_{ji}}}{\sum_{i \in F: a_{i1}=0} \prod_{j=2}^J \theta_j^{a_{ji}}} \right) > 1.
\end{aligned}$$

Therefore,

$$l(\mathbf{q}, \mathbf{p}^{(n)}) \leq l_F(\mathbf{q}_F, \mathbf{p}_F) \leq l_F(\mathbf{q}_F, \hat{\mathbf{p}}_{\mathbf{y}_F}). \quad (23)$$

Combining (22) and (23),

$$l(\mathbf{q}, \boldsymbol{\delta}^{(n)}) \leq l_F(\mathbf{q}_F, \boldsymbol{\delta}_F) \leq l_F(\mathbf{q}_F, \hat{\boldsymbol{\delta}}_{\mathbf{y}_F}), \quad (24)$$

and

$$\sup_n l(\mathbf{q}, \boldsymbol{\delta}^{(n)}) \leq l_F(\mathbf{q}_F, \hat{\boldsymbol{\delta}}_{\mathbf{y}_F}).$$

Hence, whenever $\tilde{\boldsymbol{\delta}}^{(n)} \rightarrow \tilde{\boldsymbol{\delta}}$ as $n \rightarrow \infty$, $l(\mathbf{q}, \tilde{\boldsymbol{\delta}}^{(n)}) \rightarrow l_F(\mathbf{q}_F, \hat{\boldsymbol{\delta}}_{\mathbf{y}_F})$.

Therefore, $l(\mathbf{q}, \tilde{\boldsymbol{\delta}}_{\mathbf{y}}) = \sup l(\mathbf{q}, \boldsymbol{\delta}) = l_F(\mathbf{q}_F, \hat{\boldsymbol{\delta}}_{\mathbf{y}_F})$, which concludes the proof of (ii). \square

The uniqueness claim in (iii) follows from the convexity of the log-likelihood function. The proof is similar to the one given by Lauritzen [1996, Proposition 4.7] for the case of extended log-affine models, and is thus omitted. In order to prove the second claim, suppose first that there exists a facial set F such that $\text{supp}(\mathbf{q}) \subseteq F$. Let F be the minimal of such sets. As shown in the proof of (ii), the MLE $\hat{\boldsymbol{\delta}}_{\mathbf{y}_F}$ under $RM_{\boldsymbol{\delta}_F}(\mathbf{A}_F)$ exists, and, from (8),

$$\mathbf{A}_F \hat{\boldsymbol{\delta}}_{\mathbf{y}_F} = \gamma \mathbf{A}_F \mathbf{q}_F, \text{ for some } \gamma > 0, \text{ and, if } \boldsymbol{\delta} \equiv \mathbf{p}, \mathbf{1}' \hat{\boldsymbol{\delta}}_{\mathbf{y}_F} = 1.$$

The MLE under $\overline{RM}_\delta(\mathbf{A})$ is equal to $\tilde{\boldsymbol{\delta}}_{\mathbf{y}} = (\hat{\boldsymbol{\delta}}_{\mathbf{y}_F}, \mathbf{0}_{\mathcal{I}\setminus F})$. As $\tilde{\boldsymbol{\delta}}_{\mathbf{y},F,i} = 0$ for $i \notin F$, $\mathbf{A}\tilde{\boldsymbol{\delta}}_{\mathbf{y}} = \gamma\mathbf{A}\mathbf{q}$, and, in the case of probabilities, $\mathbf{1}'\hat{\boldsymbol{\delta}}_{\mathbf{y}_F} = 1$.

If, for all facial sets F , $\text{supp}(\mathbf{q}) \not\subseteq F$, then the MLE $\tilde{\boldsymbol{\delta}}_{\mathbf{y}}$ under the extended model exists and is also the MLE under $RM_\delta(\mathbf{A})$. In this case, (8) holds and is the same as (18), which completes the proof. \square

Acknowledgments

The second author was supported in part by Grant K-106154 from the Hungarian National Scientific Research Fund (OTKA).

References

- J. Aitchison and S. D. Silvey. Maximum-likelihood estimation procedures and associated tests of significance. *J. Roy. Statist. Soc. Ser.B*, 22:154–171, 1960.
- A. H. Andersen. Multidimensional contingency tables. *Scand. J. Statist.*, 1:115–127, 1974.
- O. E. Barndorff-Nielsen. *Information and exponential families*. Wiley, New York, 1978.
- D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.
- Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice*. MIT, 1975.
- L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.
- S. Brin, R. Motwani, and R. Silverstein. Beyond market basket: generalizing association rules to correlations. In J. M. Peckham, editor, *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD 1997)*, Tucson, AZ, USA, pages 265–276, 1997.
- L. D. Brown. *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayward, Calif., 1988.
- I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.
- I. Csiszár and F. Matúš. Information projections revisited. *IEEE Trans. Inform. Theory*, 49: 1474–1490, 2003.
- R. J. Evans and A. Forcina. Two algorithms for fitting constrained marginal models. *Comput. Statist. Data Anal.*, 2013.

- S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Ann. Statist.*, 40:996–1023, 2012.
- D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34:1463–1492, 2006.
- B. Grünbaum. *Convex polytopes*. Springer, 2003.
- S. J. Haberman. *The analysis of frequency data*. The University of Chicago Press, 1974.
- A. Klimova and T. Rudas. *gIPFrm: Generalized Iterative Proportional Fitting for Relational Models*, 2014. URL <http://cran.r-project.org/web/packages/gIPFrm/index.html>. [accessed on 30 January 2015] R package version 2.0.
- A. Klimova and T. Rudas. Iterative scaling in curved exponential families. *Scand. J. Statist.*, 2015. doi: 10.1111/sjos.12139.
- A. Klimova, T. Rudas, and A. Dobra. Relational models for contingency tables. *J. Multivariate Anal.*, 104:159–173, 2012.
- S. L. Lauritzen. *Graphical models*. Oxford Statistical Science Series. Oxford University Press, 1996.
- J. Rauh, T. Kahle, and N. Ay. Support sets in exponential families and oriented matroid theory. *International Journal of Approximate Reasoning*, 52:613–626, 2011.
- A. Rinaldo. On maximum likelihood estimation in log-linear models. Technical Report 833, Carnegie Mellon Univ., 2006.
- B. Sturmfels. *Gröbner bases and convex polytopes*. AMS, Providence RI, 1996.